面向旅游在线评论情感词典构建方法 *

严仲培 [a]，陆文星 [a,b]，束　柬 [a]，王彬有 [a]

(合肥工业大学 a. 管理学院; b. 过程优化与智能决策教育部重点实验室, 合肥 230009)

摘　要：旅游在线评论情感分析的基础是情感词典的构建。在领域情感词典构建过程中，通常仅使用词频作为筛选种子词集的标准，而并未考虑其内部词语的关联程度，这会导致种子词集聚类效果不明显，进而影响情感词语归类精度。因此，基于词向量模型，提出一种情感词典种子词集筛选方法。该方法将情感词语以向量形式表征并计算词向量间距离，形成种子词集的筛选标准和分类依据，再通过类别判断形成在线评论的情感词典。最后，构建了山岳型旅游景区在线评论情感词典，并通过对比实验验证了方法的有效性，对提高情感词语归类精度和旅游在线评论情感词典的构建起到了积极的作用。

关键词：旅游在线评论；情感词典；词向量；山岳型景区

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2017.12.0812

Construction method of sentiment lexicon for online travel reviews

Yan Zhongpei[a], Lu Wengxing[a, b], Shu Jian[a], Wang Binyou[a]

(*a. School of Management*, *b. Key Laboratory of Process Optimization & Intelligent Decision-making of Ministry of Education*, *Hefei University of Technology*, *HeFei* 230009, *China*)

Abstract: The basis of emotional analysis for the online travel reviews is the construction of the sentiment lexicon. In the traditional process of constructing the field emotional dictionary, the word frequency is usually used as the criterion of screening the seed word set, instead of the association degrees of the internal words, which will lead to the effect of the seed word set clustering not that obvious, thus affecting the emotional word classification accuracy. Therefore, this paper proposed a method of seed word collection based on word vector, which expressed the emotional words in vector form and calculate the distance between word vectors as a selection criteria and classification basis of the seed word set. Finally, the emotional dictionary of the mountain scenic area was constructed, and the validity of the method is verified by a series of comparison experiments. This paper plays a positive role in improving the accuracy of emotional words and the construction of sentiment lexicon on tourism online travel reviews.

Key words: online travel reviews; sentiment lexicon; word vector; mountain scenic area

0　引言

旅游在线评论是指游客针对某一旅游景区或景点，通过专门的旅游评论网站(如 TripAdvisor、蚂蜂窝、百度旅游等)，网络社区(如百度贴吧等)或者其他渠道发表的带有自身情感或喜好的评论性文字。研究表明，旅游者获取旅游信息的渠道，除了图片之外，最重要的是网上的旅游在线评论信息[1]。随着互联网的不断发展，消费者通过网络购买商品及服务的行为不断增加，旅游已作为一种无形的、以服务为性质的产品，逐渐被列入电子商务的范围内，并在网络产品消费中占有越来越重的比例[2]。但由于消费者获取信息的不对称性导致了较大的环境不确定性，使其对旅游结果产生感知风险，进而影响了消费者的购买决策[3]。因此，如何通过游客在线评论信息分析游客情感，进而改进自身的服务水平以吸引更多旅游者，已成为旅游景区管理者需要考虑的重要问题之一。

旅游在线评论情感分析的基础是旅游在线评论情感词典。在情感词典的构建过程中，国外研究人员一般以 WordNet 为基础，进行英文情感词典的构建研究[4]。Baccianella 等人[5]基于 WordNet，构建了相关领域普遍认可的 SentiWordNet 情感词典。Turney[6]提出并改进了 PMI 算法，并在情感分类过程中取得了

收稿日期：2017-12-21；**修回日期：**2018-02-05　**基金项目：**国家自然科学基金重点项目（71331002）；国家自然科学基金青年项目（71601061）；中央高校基本科研业务费专项资金项目（JZ2015HGBZ0470，JZ2015HGBZ0468）

作者简介：严仲培（1993-），男，硕士研究生，主要研究方向为信息系统、数据挖掘、文本分析、情感挖掘等（yan201551@163.com）；陆文星（1971-），男，副教授，博士，主要研究方向为信息管理、项目管理等；束柬（1989-）男，博士研究生，主要研究方向为云计算、信息安全、信任机制等；王彬有（1990-），男，博士研究生，主要研究方向为旅游预测，评价.

较好的效果。 Araki 和 Yang 等人[7,8]分别通过改进的 SO-PMI 算法，构建了中文情感词典和日语情感词典。目前，国内研究人员主要基于 HowNet 词典进行中文情感词典的构建研究。例如，柳位平等人[9]以 HowNet 为基础，利用 TF-IDF 特征权值计算方法，构建了中文基础情感词典。除了基础情感词典，国内相关学者也针对不同领域，构建了相关的领域情感词典。例如，桂斌等人[10]提出了一种基于微博表情符号的微博情感词典构建方法，构建了微博情感词典。周咏梅等人[11]通过优化排序模型，构建了新闻评论情感词典。郭顺利等人[12]通过改进的 SO-PMI 方法，构建了图书评论的情感词典。

调查发现，面向旅游在线评论领域的情感研究很少，而山岳型旅游景区在线评论情感词典的构建工作几乎为空白。山岳型旅游景区的在线评论不同于其他的领域的评论，它所包含的情感词有其独特性，有些词语在其他领域很少使用，例如“鬼斧神工”“浑然天成”“巍峨”等。因此，其他领域的情感研究难以有效的应用于山岳型景区在线评论情感分析研究，使其具有一定的研究价值和意义。本文以“中国山岳旅游联盟”为主要研究对象，收集了相关景区在线评论数据，提出了一种旅游在线评论的情感词典构建方法，提高了旅游评论情感词语归类精度，并以山岳型旅游景区为例自动构建了其旅游在线评论情感词典，对旅游景区在线评论情感分析研究有着积极的意义。

1　基于词向量方法的旅游评论情感词典构建

1.1　思路概述

经典的专业领域情感词典构建过程一般包括情感分类、语料情感词匹配、种子词集筛选、情感词语归类四个大步骤。但其在种子词集筛选过程中，通常仅使用词频作为筛选标准，而并未考虑其内部词语的关联程度，这会导致种子词集聚类效果不明显，进而影响情感词语归类精度。

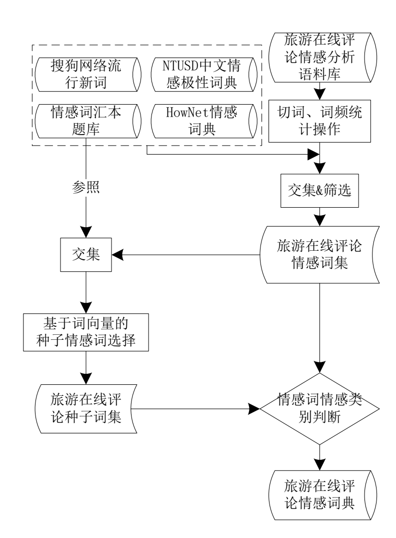


图 1　基于词向量方法的旅游在线评论情感词典构建流程

针对这一缺点，本文提出了一种基于词向量的种子情感词筛选方法。指出在筛选种子情感词的过程中，以向量形式对情感词语进行表征，并通过计算词向量之间的距离，对种子词集进行筛选，最终提高了情感词语归类的精度。本文提出的基于词向量方法的旅游在线评论情感词典构建流程大体如图 1 所示。

首先分别收集携程、蚂蜂窝等旅游网站相关在线评论数据，组成旅游在线评论情感分析语料库。利用工具进行切词、词频统计等数据清洗操作步骤后，与 HowNet 情感词典、情感词汇本题库等一系列词典组成的情感词典集进行交集操作，利用词频作为标准筛选得到旅游在线评论情感词集。随后，本文利用情感词汇本体库与上述旅游在线评论情感词集进行交集，结合本文提出的基于词向量的种子词筛选方法得到旅游在线评论种子词集。最后，通过计算情感词与种子词间的互信息进行情感词语情感类别判断，得到旅游在线评论情感词典，并以山岳型景区为例验证了该方法的有效性。

1.2　数据来源及处理

中国山岳旅游联盟是由国家旅游局倡导，国家旅游局国际司、安徽省旅游局等 11 个省的旅游局共同倡导，由黄山联合九华山、天柱山等 23 家景山岳型景区共同发起的中国第一个山岳型旅游联盟。联盟创建的主体思想是“平台共建、品牌共创、价值共享”，即在追求包容性发展、可持续发展的同时，深化科学研究与交流，推进资源与环境保护，不断彰显我国山岳历史与文化。

本文以中国山岳旅游联盟组成成员作为数据收集参照，通过图 2 对比了各大旅游网站近三年内百度指数搜索数据。

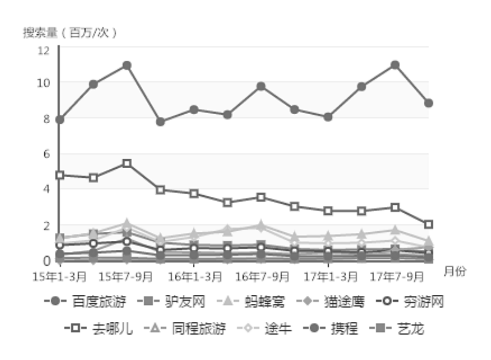


图 2　旅游网站百度指数对比图

通过图 2 可以看出，携程旅游网站搜索量在各大旅游网站中处于领先地位，其次是去哪儿以及蚂蜂窝等。但在实际操作过程中发现，一些网站如去哪儿等，网站系统默认好评过多，有效数据集过少；一些网站如途牛、同程旅游等，景点评论数据量较少或没有。为保证实验数据的有效性，本文将上述网站剔除并最终选取携程、蚂蜂窝以及百度旅游三者作为数据来源网站，利用八爪鱼数据收集软件采集了相关旅游评论数据，整合得到旅游在线评论情感分析语料库。表 1 显示了部分数据集示例。

由于某些较短的网络评论如“来过”，“很好”等，仅包含一个词语，无法具体表达个人情感因素，为防止其在后续数据处理时产生无法筛选出情感词等情况，本文对上述数据集进行了筛选。

表 1　数据集示例

景点名称	在线评论
大别山	大别山地处鄂豫皖交界处，以前交通不是很方便。但是合肥到武汉的高铁通车以后，到大别山方便了很多。可以坐动车到金寨后，再坐大巴到大别山就方便很多了。从金寨车站到大别山的景区，约 2 个多小时的车程就能到了。大别山的经典景点是天堂寨。从山下的入口一路上行，沿途瀑布、溪流，景色非常优美。到山顶就到了鄂豫的分界线了。在这里可以玩一把一脚跨两省的游戏。从这里再往前，就是湖北省的地界了。所以，也可以从湖北的罗田县上山。天堂寨景区门票 115 元，另外还要买游览车票 28 元。上下山游览约需 4 个小时。
峨眉山	10 月 7 日去的乐山，峨眉，错过了高峰，很好。峨眉山的金顶是一定要去看日出和云海的，不然意义就不是很大，如果要看日出就要穿很厚的衣服，下面有租衣服的，猴区是真的没意思去，里面的猴子都是当地人驯服了的，完全是抢劫，当地人教猴子爬到你身上，然后你就买猴粮，那些猴子很厉害，最好是在另外一条路上看，买那个竹竿就没什么必要了，因为那些猴子根本就不怕，千万不要去摸猴子，有时它要咬人，本文当时去爬金顶时，就有人被咬了，到时就要去打针了，金顶的猴子很可爱，他们就只是要吃了，只要你不去摸它就好了。
黄山	黄山是最秀美的，建议大家千万不要跟旅行团去，黄山最好玩两天。在山上住一晚，山顶的星空要比城市里的璀璨的多。黄山有比较多的奇山异石，所以，一定要看景点介绍，当然，也可以发挥你的想象力。还有，第二天一定要在早上或下午看朝霞或夕阳，建议坐缆车，注意，由于缆车落差较大，所以在乘坐的时候一定要时不时地张嘴，打哈欠，或咽吐沫，以保证耳膜不受伤害。祝大家旅行愉快。
崂山	崂山很棒，也是青岛一个很有特色的景点，因为青岛的山都是石头山，崂山风景可以和云南石林一拼啦，而且崂山是纯天然的，崂山有几个景点，我觉得最棒的就是仰口，一定要去山顶，无论你是坐缆车或者自己爬，因为到了山顶你肯定不会后悔的！其中有个山洞也是很特别的，路很崎岖，很黑，而且据说胖子过不去哦～

根据吴云芳等人的研究结果，新闻评论句长通常在 30~40 个字左右，由于旅游评论跟新闻评论都隶属于评论的一种，故在此删除评论长度小于 30 个字符以及仅包含图片评论的语料数据，最终得到 10 万条旅游评论数据作为旅游在线评论情感词典构建语料库 DataSet。接下来，本文利用 Python 编程软件对 DataSet 进行切词和词频统计操作，综合考虑语料库规模以及其他非情感词语因素造成得影响，本文选取 100 作为词频筛选标准，去除词频低于此标准的词语，最终得到山岳型景区旅游在线评论词集 M，共 2369 个词语。

由于旅游在线评论中包含了许多专业用语以及一些网络新词，现有的词典都不能完全涵盖所有的情感词。因此，本文首先收集了大连理工大学情感词汇本体库，台湾大学 NTUSD 中文情感极性词典，知网 HowNet 中文词库以及搜狗网络流行新词词典共四个现有的情感词典，整合并筛选去除重复的情感词语，最终得到旅游在线评论合并情感词典 N，共 37835 个词语。将 M 和 N 做对比，取两者交集后，以词频为依据结合哈工大同义词词林筛选去除部分重复词语，如“费力”和“吃力”等，最终形成旅游在线评论情感词集 TravelWordSet，共包含 758 个情感词。

1.3　基于词向量的种子情感词选择方法

旅游在线评论情感种子词指的是，具有强烈情感倾向性的代表性词语。本文拟采用郭顺利人提出的改进的 SO-PMI 算法[12]进行词语情感倾向性的判断，所以要针对上一步生成的 TravelWordSet 词集进行进一步筛选操作，找出旅游在线评论种子词集。对于旅游景区的在线评论，不同的评论可能包含着不同的情感，因此需要事先对其进行情感分类操作。到目前为止，心理学领域对于情感分类仍然没有一个统一的标准。一般情况下，研究者们会将用户的情感分为 4，8，10 乃至 20 类不等。本文参照徐琳宏等人的情感词汇本体库[17]构建方法，根据已有的资源，将用户在线评论情感分为乐，好，怒，哀，惧，恶，惊七大类别。

情感分类是种子词集筛选以及情感词语归类的基础。对于种子词集筛选，一般的方法是以情感词汇本体 WordSet 为基准，根据词频的大小筛选种子情感词。这种方法虽然简便，但并没有考虑词语之间的相互关系，这会导致种子词集聚类效果不明显，影响情感词语归类精度，进而影响文本情感的判别。

针对这一现象，本文采用一种基于词向量的情感词典种子词集筛选方法，指出将情感词以向量形式表征，通过计算词向量之间的距离对种子情感词集进行筛选，旨在提高种子词集的内聚程度，进而提高情感词语归类精度。Word2vec 是 Mikolov 在 2013 年提出的一种将词语表征为实数值向量的高效工具[13]，其主要利用深度学习的思想，通过 skip-grams 或 continuous-bag-of-words（CBOW）算法进行词语嵌入，并最终通过机器学习的方法探究给定文档中词语之间的语义关系。Word2vec 由于其高效性等特点，常被用来做很多 NLP 相关的工作，比如分类[14,15]、同义词查找、语义分析[16]等等。

基于上述优点，本文首先利用此工具对情感词语进行了向量形式的表示，接下来本文计算了各个词语之间的余弦距离，并借鉴萤火虫算法相关概念，将单个情感词语和其他情感词语的平均距离定义为该情感词语的情感吸引度 EAD(Emotional Attractive Degree)，其计算公式如下：

$$EAD(word_{ij}) = \frac{1}{j}\sum_{n=1}^{j} Dis(word_{ij}, word_{in}) \qquad (1)$$

其中：$word_{ij}$ 表示第 i 类情感的第 j 个情感，Dis(wordA,wordB)

代表 wordA 和 wordB 两个词向量之间的余弦距离。为了对情感类别中的情感词语进行筛选，以便为后续步骤提供种子词语，需要定义一个筛选的标准距离。在此，本文使用如下式(2)对此距离进行计算：

$$StaEAD_i = \frac{1}{j}\sum_{m=1}^{j} EAD(word_{im}) \tag{2}$$

其中 $StaEAD_i$ 表示第 i 类情感的标准筛选距离。通过以上公式，分别计算每个种子词的情感吸引度 $EAD(word_{ij})$以及每类种子词集的标准筛选距离 $StaEAD_i$，筛选剔除其中 EAD 值小于该类 StaEAD 值的词语，最终得到旅游在线评论情感词种子词集 SeedWordSet。上述具体流程如图 3 所示。

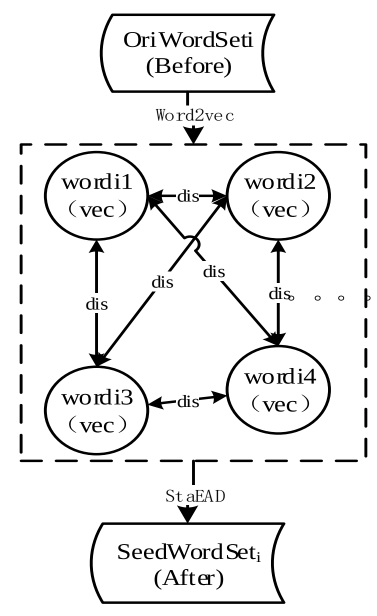


图 3　基于词向量方法的单个种子词集筛选过程

为了方便观察，本文对上述方法进行了算法形式的描述：

输入：旅游在线评论情感词集 TravelWordSet，情感词汇本体词典 WordSet。

输出：旅游在线评论情感词种子词集 SeedWordSet。

a)对于 TravelWordSet 内每一个词语 w_i，如果 w_i 属于 WordSet，则将（w_i，e_i）加入到临时情感词典 TSD 内。其中 e_i 代表每个情感词在 WordSet 中对应的情感分类，具体包括乐，好，怒，哀，惧，恶，惊七大类别。

b)对于上步骤得到的 TSD，根据每个 w_i 的 e_i，将所有情感词语分为七大情感词集，并分别记为 $OriWordSet_i$ ($1\le i \le 7$)。

c)对于每一个 $OriWordSet_i$，分别计算其中每个词的 EAD，并通过计算 StaEAD 对情感词语进行筛选，最终得到每个情感的种子词集 $SeedWordSet_i$ ($1\le i \le 7$)。

e)合并所有的种子词集，得到旅游在线评论种子词集 SeedWordSet。

f)算法输出并结束。

1.4　旅游情感词汇情感类别判断

对于词语情感类别判断方法，主要有基于知网 HowNet 的语义相似度判别方法[18]和基于词语互信息的 SO-PMI 情感词倾向性计算方法[19]。因为旅游在线评论包含的许多网络新词如“给力”“坑爹”等，在 HowNet 中无法找到其义原，无法计算语义相似度，所以不能用这种方法判断其情感类别。本文根据郭顺义等人提出的改进的 SO-PMI 算法[12]，通过计算词语和种子词集之间的互信息来判断未知词语的情感类别，实验证明了该方法的有效性。

记旅游在线评论情感词集 TravelWordSet 去除种子词集 SeedWordSet 后形成的需要判断的旅游在线评论情感词集为 TravelWordSetX。针对 TravelWordSetX 里面的每一个词语 $word_x$，计算其与每一类情感的种子词集 $SeedWordSet_i$ 的关联度 $SO_PMI(word_x,\ SeedWordSet_i)$，如式(3)所示。

$$SO_PMI(word_x, SeedWordSet_i) = \sum_{w_{ij}\in SeedWordSet_i} \log_2\left(\frac{N\times hit(word_x, w_{ij})}{M_i\times d\times hit(word_x)\times hit(w_{ij})}\right) \tag{3}$$

其中：N 代表语料库中所有词语的总次数，hit 表示词语的词频数，w_{ij} 表示单个情感种子词集 $SeedWordSet_i$ 中的每一个词语，$hit(word_x,\ w_{ij})$表示词语 $word_x$ 和 w_{ij} 在同一条旅游在线评论中的共现次数。d 表示两个词语之间的共现距离，Mi 表示单个种子词集中情感词的个数。最终，通过对比每个词语的相对于种子词集的 SO_PMI 大小，将 TravelWordSetX 里面的词语进行情感分类，整理后得到旅游在线评论情感词典。

2　实验分析

为了验证基于词向量的旅游在线评论情感词典构建方法的有效性进，本文设计了四组对比实验，分别对该方法运用于旅游评论情感词语情感类别判定的适用性和准确性进行了对比评估。

首先，根据网络上获取的中国山岳旅游联盟旅游在线评论数据，利用本文提出的基于词向量的情感词典构建方法得到相应的种子词集并记其为 SeedWordSet1，共包含 162 个情感词，如表 2 所示。

接下来，本文仅利用词频作为依据对种子词集进行人工判别筛选，并记此方法产生的种子词集为 SeedWordSet2，共有 158 个情感词，如表 3 所示。

由表 2 和 3 对比可以看出，本文提出的基于词向量方法的种子词筛选方法与仅利用词频方法筛选得到的种子词集有所不同。如，情感类别为“好”里面的情感原始种子词“不错”一词，通过词向量方法计算出其与原始词集内部词语间的平均关联度为 0.264，远远小于本文进行情感种子词筛选的标准 StaEAD 计算出来的值 0.53，说明其与种子词集内部的词语关联性不大，这将导致后续情感词类别判断过程中 SO-PMI 值计算较小，影响情感词语判别的精度，故对其进行剔除操作。

利用 2.2 节中的旅游在线评论情感词集 TravelWordSet 去除上述两个种子词集中出现的词语，可得测试词集 TestWordSet，总共 534 个情感词。

表 2　旅游在线评论种子词集

情感类别	情感种子词
乐(22 个)	惬意,心旷神怡,放松,静谧,快乐,幸福,豁然开朗,尽情,高兴,舒畅,曲径通幽,神清气爽,自豪,愉悦,诗情画意,水天一色,天上人间,飘飘欲仙,赏心悦目,不亦乐乎,沁人心脾,悠然自得
好(45 个)	鬼斧神工,雄伟,优美,美不胜收,奇观,名不虚传,山清水秀,壮丽,丰富,迷人,完美,天下第一,郁郁葱葱,巍峨,流连忘返,高大,风景如画,挺拔,赞叹,美妙,惟妙惟肖,闻名,神圣,名副其实,金碧辉煌,气势磅礴,和谐,清秀,陶醉,湖光山色,留恋,瑰宝,大好河山,葱郁,沉醉,流连,云蒸霞蔚,崇拜,引人入胜,浑然天成,赏心悦目,神往,如痴如醉,波光粼粼,情不自禁
怒(3 个)	过分,爆发,失落
哀(20 个)	遗憾,失望,后悔,无奈,一场空,绝望,不幸,悲剧,无语,孤独,冷清,伤心,美中不足,孤单,荒凉,破败,寂寞,大失所望,萧条,倒霉
惧(17 个)	惊险,害怕,悬崖峭壁,吓人,恐怖,困难,可怕,小心翼翼,漆黑,体力不支,恐惧,艰险,心惊胆战,惊恐,望而却步,胆怯,望而生畏
恶(40 个)	拥挤,没意思,浪费,无聊,麻烦,忽悠,不怎么样,勉强,骗人,崩溃,难受,难吃,怀疑,嘈杂,单调,生气,糟糕,尴尬,恶心,压抑,浮躁,人头攒动,恨不得,吵闹,恶劣,灰蒙蒙,喧哗,狼狈,鄙视,枯燥,扫兴,差劲,辜负,莫名其妙,折磨,未必,名不副实,吃不消,欺骗,可恶
惊(15 个)	奇特,神秘,世外桃源,惊叹,奇妙,叹为观止,奇迹,罕见,惊奇,奇异,不可思议,惊人,震撼人心,意想不到,千奇百怪

表 3　仅依据词频筛选山岳型景区在线评论种子词集

情感类别	情感种子词
乐(19 个)	自然,舒服,享受,宁静,惬意,心旷神怡,天堂,豁然开朗,曲径通幽,神清气爽,恬静,诗情画意,水天一色,赏心悦目,不亦乐乎,沁人心脾,悠然自得,兴致勃勃,风和日丽
好(45 个)	不错,值得,喜欢,漂亮,壮观,美丽,著名,仙境,圣地,安静,天然,秀丽,鬼斧神工,雄伟,气势,美不胜收,奇观,名不虚传,山清水秀,壮丽,迷人,天下第一,郁郁葱葱,巍峨,流连忘返,青山绿水,高大,屹立,风景如画,赞叹,惟妙惟肖,胜地,名副其实,金碧辉煌,古色古香,气势磅礴,苍翠,奇秀,陶醉,湖光山色,瑰宝,大好河山,沉醉,慕名,喜爱
怒(3 个)	过分,爆发,失落
哀(20 个)	可惜,遗憾,失望,后悔,无奈,艰辛,痛苦,怀念,一场空,绝望,不幸,悲剧,无语,冷清,孤零零,伤心,美中不足,破碎,孤单,荒凉
惧(17 个)	震撼,小心,惊险,害怕,悬崖峭壁,险要,恐怖,小心翼翼,漆黑,体力不支,地狱,恐惧,艰险,胆战心惊,生怕,阴森
恶(40 个)	不好,担心,逃票,严重,拥挤,没意思,吃力,费劲,好不容易,疲惫,浪费,无聊,麻烦,忽悠,罢了,不怎么样,不及,缺点,不过如此,勉强,骗人,勉强,喧闹,崩溃,污染,难吃,嘈杂,单调,糟糕,离谱,闷热,人头攒动,恨不得,岂有此理,不合理,扫兴,差劲,辜负,莫名其妙,折磨
惊(15 个)	神奇,奇特,神秘,世外桃源,惊叹,奇妙,叹为观止,奇迹,罕见,惊奇,不可思议,震撼人心,意想不到,从天而降,千奇百怪

对于测试词集，首先采用人工判别的方法，选取五位参与人员，分别对每个情感词进行人工判别情感分类，最终通过统计五个人的判别结果，对比讨论得出人工判别结果作为情感词情感倾向标准值。接下来，采用实验组 2 和 4 分别对 SeedWordSet1 和 SeedWordSet2 使用 2.4 节提出的改进的 SO-PMI 方法对 TestWordSet 中词语情感倾向性进行判别。为了验证 2.4 节方法的适用性，采用实验组 1 和 3 分别对 SeedWordSet1 和 SeedWordSet2 使用原版的 SO-PMI 方法进行词语倾向性判别作为对比实验，最终得出的结果如下表 4 所示。为了方便观察，本文利用软件对算法判别结果进行了表示，如图 4 所示。

表 4　对比实验结果

情感分类	乐	好	怒	哀	惧	恶	惊	总正确率/%
人工判别组	72	143	28	68	56	126	45	
实验组 1	64	124	34	48	62	128	74	74.5
实验组 1 正确判别	56	107	16	34	42	109	34	
实验组 2	75	134	27	56	52	134	56	83.5
实验组 2 正确判别	63	125	13	47	48	112	38	
实验组 3	59	136	26	53	53	143	64	68.7
实验组 3 正确判别	48	103	14	28	40	103	31	
实验组 4	68	131	38	55	64	126	52	77.9
实验组 4 正确判别	58	112	14	42	45	109	36	

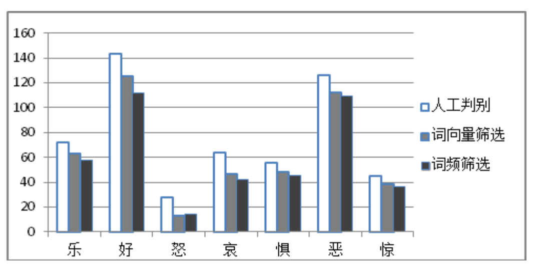


图 4　算法判别结果对比图

实验结果表明，本文采用的改进的 SO-PMI 方法总体上较原版的 SO-PMI 算法正确率高，故改进的 SO-PMI 算法可以利用于旅游在线评论情感词情感类别判断过程中。另一方面从结果可以看出，种子词的数量对情感词类别判断仍存在一定的影响，如“怒”的种子词较少，故其准确率，召回率等相对值都较低，影响了总体的均值。另外，一些词语如“精神”等，可以作为名词也可以作为形容词，也可能会影响总体的判断精度。

接下来，为了评估本文提出的基于词向量的种子词集筛选方法的算法性能，本文采用了准确率（P 值），召回率（R 值）以及 F 值三项作为衡量指标，对算法进行评估，计算后的结果如表 5 所示。

由表 4 可以看出，利用词向量筛选后的 SeedWordSet1 进行情感词情感判别准确率平均值为 0.79，召回率平均值为 0.79，F 值平均值为 0.79，均高于仅利用词频筛选出的 SeedWordSet2。由此可得，本文提出的利用词向量方法的旅游在线评论情感词

集种子词筛选方法对于旅游情感词典构建具有较高的准确性和可利用性。

表 5 基于词向量的种子词集筛选算法评估

情感分类		乐	好	怒	哀	惧	恶	惊	平均
SeedWordSet1	P	0.88	0.87	0.46	0.46	0.73	0.86	0.89	0.79
改进的 SO-	R	0.84	0.93	0.48	0.84	0.92	0.84	0.68	0.79
PMI 方法	F	0.86	0.90	0.47	0.78	0.89	0.86	0.75	0.79
SeedWordSet2	P	0.81	0.78	0.50	0.66	0.80	0.87	0.80	0.74
改进的 SO-	R	0.85	0.85	0.37	0.76	0.70	0.87	0.69	0.73
PMI 方法	F	0.83	0.82	0.42	0.71	0.75	0.87	0.74	0.73

3 结束语

本文提出了一种基于词向量的旅游在线评论情感词典构建方法，指出在旅游在线评论情感词典构建种子词筛选过程中，利用 Word2Vec 工具将情感词表示成为向量形式，并通过向量之间的距离对情感词典种子词集进行筛选，最终通过对情感词情感类别进行判断，得到旅游在线评论情感词典。并以山岳型景区为例，验证了该方法的有效性。本文中基面向山岳型景区的在线评论情感分析方法具有一定的普适性，可为其他领域情感词典的构建提供一定的参考价值。

本文研究中虽取得一定阶段性成果，但仍存在部分不足：如语料数量仍然不够庞大，造成最终一类情感种子词集较少，影响了判别的精度。拟在下一步研究中做到以下改进计划：从其他旅游评论网站以及外网爬取相关评论数据，扩大研究语料库并将方法向其他非山岳型旅游景区进行推广，提高该旅游景区在线评论情感词典构建方法的普适性。

参考文献：

[1] 胡兴报，苏勤．黄山国内旅游者网络旅游信息搜索行为研究 [J]．安徽师范大学学报：自科版, 2011, 34 (3): 282-287.

[2] 中国互联网络信息中心（CNNIC）. 2016 年第 38 次《中国互联网络发展状况统计报告》. 2016

[3] 许晖，许守任，王睿智．消费者旅游感知风险维度识别及差异分析 [J]．旅游学刊, 2013, 28 (12): 71-80.

[4] Andreevskaia A, Bergler S. Mining WordNet for a fuzzy sentiment: sentiment tag extraction from wordnet glosses [C]// Proc of Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference. 2006.

[5] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3. 0: an enhanced lexical resource for sentiment analysis and opinion mining [C]// Proc of International Conference on Language Resources and Evaluation. 2010: 83-90.

[6] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews [C]// Proc of Annual Meeting of the Association for Computational Linguistics. 2002: 417-424.

[7] Wang G, Araki K. Modifying SO-PMI for Japanese Weblog opinion mining by using a balancing factor and detecting neutral expressions [C]// Proc of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. 2007: 189-192.

[8] Yang A M, Lin J H, Zhou Y M, *et al.* Research on Building a Chinese Sentiment Lexicon Based on SO-PMI [J]. Applied Mechanics & Materials, 2013, 263-266: 1688-1693.

[9] 柳位平，朱艳辉，栗春亮，等．中文基础情感词词典构建方法研究 [J]．计算机应用, 2009, 29 (10): 2875-2877.

[10] 桂斌，杨小平，张中夏，等．基于微博表情符号的情感词典构建研究 [J]．北京理工大学学报, 2014 (5): 537-541.

[11] 周咏梅，阳爱民，杨佳能．一种新闻评论情感词典的构建方法 [J]．计算机科学, 2014, 41 (8): 67-69.

[12] 郭顺利，张向先．面向中文图书评论的情感词典构建方法研究 [J]．现代图书情报技术, 2016, 32 (2): 67-74.

[13] Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv: 1301. 3781, 2013.

[14] Giatsoglou M, Vozalis M G, Diamantaras K, *et al.* Sentiment analysis leveraging emotions and word embeddings [J]. Expert Systems with Applications, 2017, 69: 214-224.

[15] Yuan Y, He L, Peng L, *et al.* A new study based on Word2vec and cluster for document categorization [J]. Journal of Computational Information Systems, 2014, 10 (21): 9301-9308.

[16] Xue B, Fu C, Zhan S. A study on sentiment computing and classification of Sina Weibo with word2vec [C]// Proc of IEEE International Congress on Big Data. 2014: 358-363.

[17] 徐琳宏，林鸿飞，潘宇，等．情感词汇本体的构造 [J]．情报学报, 2008, 27 (2): 180-185.

[18] 朱嫣岚，闵锦，周雅倩，等．基于 HowNet 的词汇语义倾向计算 [J]．中文信息学报, 2006, 20 (1): 16-22.

[19] 杜锐．面向中文微博文本的情感分类研究 [D]．株洲：湖南工业大学, 2014.